GENERALIZED LOG-LOGISTIC MODEL FOR ANALYSIS OF ENVIRONMENTAL POLLUTANT DATA

Karan P. Singh, Warsono, Alfred A. Bartolucci
Department of Biostatistics, School of Public Health
University of Alabama at Birmingham
Birmingham, AL 35294, U. S. A

Abstract: Environmental pollution studies conducted to monitor ambient levels and to quantify the concentration of various pollutants entering a given environmental area are of great interest for possible adverse-health effects. Of particular importance in environmental data analysis is to select appropriate probability models. The previous studies indicate that none of the probability models, including the classical lognormal, has been identified to be superior to others in a general sense. To address this problem, the purpose of this paper is twofold. Firstly, we introduce a generalized log-logistics distribution as a general model in fitting environmental pollutant data. The family of the generalized log-logistics distribution include several well-known distributions in modeling data of environmental pollutant concentrations, such as lognormal, Weibull, and gamma as special cases. Secondly, by applying the proposed model to some environmental data sets, we explore the possibilities of using this model as a general probability model for fitting environmental-quality data.

1. Introduction

Environmental pollution studies conducted to monitor ambient levels and to quantify the concentration of various pollutants entering a given environmental area are of great interest for possible adverse-health effects. Selecting appropriate probability models for the data is an important step in environmental data analysis. These probability models may become the basis for estimating the parameters to meet the evolving information needs of environmental quality management. Unfortunately, the environmental pollution data are frequently skewed to the right; that is, they have a long tail toward high concentration. Therefore, the validity of applying the normal distribution for curve fitting of these type of data may be questioned. One way of modeling this type of distribution is to find a transformation of data values so that the transformed values conform more closely to the normal distribution, and the logarithmic transformation is often applied in this context to pollution data. However, parameter estimates of the transformed data are rarely of interest. The estimate of the mean, for example, in the original scale of measurement is the primary purpose of environmental study.

A further complication is introduced by the fact that there are a number of observations measured as less than detection limit (DL) established by analytical laboratories. The analysts may report them as nondetect (ND) or less than detection limit (LDL) rather than numerical values. Even if the data are normally distributed, the presence of left-censoring creates some difficulties when applying classical methods because one will be uncertain as to what to use for censored values. In practical applications, to handle the censored data, many analysts ignore the values of observations below the DL or set them equal to zero, the DL or the DL divided by two (DL/2) prior to parameters estimation. Replacing with the DL/2 implicitly assumes a uniform distribution between zero and the DL. But the deletion or the replacement gives biased estimates of the parameters, and the intensity of the bias will be worse as the degree of censoring increases (Newman et al., 1989). Newman et al. (1989) do not recommend the use of such techniques.

Although the lognormal distribution has been widely employed to represent pollution concentration data, a fact that also should be pointed out is that it is possible that other distributions might work better. The lognormal distribution with different parameters is sometimes appropriate. Dealing with air-quality data, Larsen (1974) added a third parameter, an increment, to the lognormal distribution. The third parameter is either a positive or negative increment that is added to every observed concentration until a curved log-probability plot is transformed into a fairly straight line. Mage and Ott (1978) called their model the censored three-parameter lognormal model. Mage and Ott do not suggest the automatic use of a particular model, because failure to consider the validity of the model, if hypothesis tests are involved, can lead to predictions that are not supported under scientific scrutiny.

Berger, Melice, and Demuth (1982) examined the goodness-of-fit based on the extreme values and the median in fitting a gamma distribution to daily atmospheric sulfur dioxide (SO2) concentrations in the Gent region of Belgium. They found that the gamma distribution provided a better representation of the whole ensemble than the usual lognormal. Jakeman and Taylor (1985) also observed that gamma models provide a better representation of acid-gas concentrations in an industrial airshed than does the lognormal model.

In published literatures, none of the probability models, including the classical lognormal, has been identified to be superior to others in a general sense. Among the general models, the generalized log-logistic (GLL) distribution has good potential for fitting environmental pollutant data. The GLL distribution is an extension of the log-logistic distribution. The log-logistic distribution is similar in shape to the lognormal distribution, but it may be more convenient to apply. This is because of its greater mathematical simplicity, especially when dealing with the censored data, Singh (1989) and Singh et. al. (1994).

One approach to determining an appropriate model is to use a very general model that includes a suitable model as a special case. Although in environmental studies the GLL distribution is a relatively "unknown" distribution, as mentioned earlier the skewness and the heavy tail of the GLL distributions seem to make it suitable for modeling environmental pollution data. Also, the family of the GLL distribution is quite rich and includes a number of submodels that are very common distributions in fitting pollutant concentration data. Therefore, the GLL distribution has desirable features and seems to be a promising distribution for environmental modeling. Thus, in this paper we propose to consider the use of the family of GLL distributions in fitting pollutant concentration data. The family may provide more flexibility to fit environmental data when the skewness, kurtosis, or other moments of the distribution fail to conform to lognormality. Thus, the family of GLL distributions may become a good alternative to the lognormal distribution. The overall objective is to provide analysts, especially those who work in environmental areas, more latitude in selecting various models.

By applying the proposed model to various sets of data, we explore the possibility of using GLL distribution as a general probability model for representing environmental quality data. For comparison purposes , we also consider the three-parameter GLL distribution where m_1 = m, and m_2 = 1, denoted by GLL(m,1); the three-parameter GLL distribution where m_1 = 1, and m_2 = m, denoted by GLL(1,m); the log-logistic distribution, denoted by GLL(1,1); and lognormal distribution.

2. Log-Normal Distribution

The random variable X is said to have a two-parameter lognormal (LN) distribution if the random variable $Y = \ln X$, where $0<X$, is normally distributed with mean μ and variance σ^2. The probability density function (PDF) of X is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln x-\mu)^2} , \quad 0<x.$$

The cumulative distribution function (CDF) of the lognormal distribution is

$$F(x) = \Phi\left(\frac{\ln x-\mu}{\sigma}\right)$$

The LN distribution has a long history of application in the field of environmental pollution. A rich literature has been published over the past two decades, suggesting that pollutant concentration data tend to be lognormally distributed. The decision to apply the LN distribution in fitting pollutant concentration data can be attributed to the work of Larsen (1969, 1973, 1974). Using graphical techniques, he concluded that regardless of pollutant, city, or averaging time, the air concentration distributions are approximately lognormally distributed. An excellent review of the history of the applications of probability models, especially lognormal models, to aerometric data is given by Mage (1981). Under lognormally assumption, El-Shaarawi (1989) examined several methods for making inferences about the levels of many metals and organic contaminants in ambient water samples from the Niagara River. More recently, the applications of the lognormal model to air-, soil-, and water-quality data are presented in considerable detail by Ott (1995).

Having developed physical mechanisms generating environmental quality data, Ott (1995) provided an argument as to why the LN distribution is so ubiquitous in environmental phenomena. The LN distribution has been fitted not only for air quality data, but also for water quality and geological data. Ott's explanations involve the central limit theorem and the diffusion law.

A number of investigators also considered other distributional forms for environmental-quality data. Using the sum-of-squares error as the goodness-of-fit criterion, Bencala and Seinfeld (1976) showed that Weibull models produce lower values than that of the lognormal model for five of eight CO data sets. But they stated that the LN model is convenient from a practical point of view. A similar study comparing the LN model with other probability models is also carried out by other researchers, for example, by Berger et al. (1982), Simpson, Butt, and Jakeman (1984), Jakeman and Taylor (1985), and Taylor et al. (1986). Georgopoulos and Seinfeld (1982) presented a critical review of statistical distributions, such as Weibull, gamma, and many others, and stated that the LN distribution has been the most popular in representing urban air pollutant concentration data.

For more detail, the reader is referred to Warsono, Singh and Bartolucci (1996).

3. Generalized Log-Logistic Distribution

Singh (1989) suggested a generalized log-logistic (GLL) distribution, which is a natural extension of the log-logistic (LL) distribution in modeling data of lung and other cancers. He also demonstrated the flexibility of the GLL model in fitting lung cancer survival data. Further illustrations of the GLL application in modeling breast cancer survival data are given by Singh et al. (1994).

Let a random variable X have four-parameter GLL distribution with shape parameters m_1 and m_2, denoted by $X \sim$ GLL(m_1,m_2).

Let $B(m_1,m_2)$ be the complete beta function, which is defined as follows

$$B(m_1,m_2) = \frac{\Gamma(m_1)\ \Gamma(m_2)}{\Gamma(m_1+m_2)}$$

where Γ is the gamma function, and

$$F(x) = [1 + e^{-[\beta+\alpha\ln(x)]}]^{-1}$$

be the log-logistic distribution function. Then after simplification, the PDF of the GLL(m_1,m_2) distribution is given by

$$g(x) = \frac{\alpha}{xB(m_1,m_2)}\ [F(x)]^{m_1}\ [1-F(x)]^{m_2}.$$

For more detail, the reader is referred to Warsono, Singh and Bartolucci (1996).

4. Applications of Models To Data Sets - Examples

The first example is of uncensored data of mercury concentration in ppm in 115 sample swordfish published by Lee and Krutchkoff (1980). The maximum-likelihood estimates of the parameters obtained by the GLL(m_1,m_2) model are $\hat{\alpha}$ = 14.0428, β = -5.4922, $\hat{m}_1$ = 0.1192, and $\hat{m}_2$ = 0.4342. The 95% asymptotic confidence intervals for α, β, m_1, and m_2 are [13.7578,14.3278], [-5.6749,-5.3095], [0.1167,0.1217], and [0.4094,0.4580], respectively. The GLL(m,m) fit yields $\hat{\alpha}$ = 17.5504, β = -1.7498, and $\hat{m}$ = 0.1303. The 95% asymptotic confidence intervals for α, β, and m are [17.1572,17.9436], [-1.9612,-1.5384], and [0.1257,0.1349], respectively. Given in Table 1 are the values of log-likelihood functions, and Akaike Information Criterion (AIC) for GLL(m_1,m_2), GLL(m,m), GLL(m,1), GLL(1,m), GLL(1,1) and lognormal distributions.

From Table 1, it is clear that the value of the log-likelihood for the GLL(m_1,m_2) distribution is considerably larger than those using the lognormal and GLL(1,1) distributions and slightly larger than those using the GLL(m,1), GLL(1,m), and GLL(m,m) distributions. Note the values using the GLL(m,1), GLL(m,1), and GLL(m,m) distributions are remarkably larger than that of the traditional lognormal distribution and of the GLL(1,1) distribution is slightly larger than that of the lognormal distribution. Therefore, by looking of the maximum log-likelihood values, the GLL(m_1,m_2) distribution seems to be a better statistical model in fitting the data of mercury concentration. Moreover, the AIC of the GLL(m_1,m_2) and GLL(m,1) distributions are considerably lower than those of log-logistic, GLL(1,m), and GLL(m,m) distributions. Consequently, from the AIC value standpoint, the GLL(m_1,m_2) and GLL(m,1) distributions may provide better description of the data of the mercury concentration.

Figures 1-3 present graphs of the fitted CDFs of models superimposed on the empirical distribution function. The graphs suggest that there is improvement in fit using the GLL distributions. In particular, the GLL(m_1,m_2) performs considerably better than the lognormal, GLL(1,1), GLL(1,m), & GLL(m,m) distributions and slightly better than the GLL(m,1) distribution. Notice that the other GLL distributions also appear to fit better than the classical lognormal distribution.

Table 1: Values of the log-likelihood functions and of AIC for models fitted to mercury data

Model	Log-Likelihood	AIC
GLL(m,1)	- 81.1847	168.3694
GLL(1,m)	- 85.5972	177.1974
GLL(m,m)	- 933.9988	193.9976
GLL(m_1,m_2)	- 80.9181	169.8362
Lognormal	- 114.1681	-

The second example uses the data on copper concentrations with 34 uncensored and 14 censored observations in the San Joaquin Valley, California, published by Millard and Deverel (1988). The maximum-likelihood estimates of the parameters obtained using GLL(m_1,m_2) model are $\hat{\alpha}$ = 0.6560, β = 1.8748, $\hat{m}_1$ = 41.9214, and $\hat{m}_2$ = 3.6089. The 95% asymptotic confidence intervals for α, β, m_1, and m_2 are [0.6551,0.6569], [1.8739,1.8758], [41.5963, 42.2465], and [3.5251,3.6927], respectively. The GLL(m,m) fit yields $\hat{\alpha}$ = 0.1103, β =-0.1184, and $\hat{m}$ = 199.5571. The 95% asymptotic confidence intervals for of α, β, and m are [0.11029,0.11031], [-0.707,0.4702], and [198.4403,200.6739], respectively. Table 2 contains values of log-likelihood function and AIC under GLL(m_1,m_2), GLL(m,m), GLL(m,1), GLL(1,m), and GLL(1,1), log-logistic distributions, and the lognormal distribution.

From Table 2, it can be seen that the value of log-likelihood of GLL(m_1,m_2) distribution is larger than those of the lognormal, GLL(1,1), GLL(m,1), GLL(1,m), and GLL(m,m) distributions. Also note that these values of other GLL distributions are larger than that of the lognormal distribution. Hence, the GLL(m_1,m_2) model seems to be a promising model in fitting data of copper concentration. However, the AIC of the GLL(1,1) distribution is slightly lower than the compared distribution. Thus, in this case, because of mathematical simplicity, the log-logistic distribution is preferable over the three-parameter and four-parameter GLL distributions.

Table 2: Values of the log-likelihood functions and of AIC for models fitted to copper data

Model	Log-Likelihood	AIC
GLL(m,1)	- 96.4600	239.1120
GLL(1,m)	- 96.9296	239.9120
GLL(m,m)	- 96.3422	238.5548
GLL(m_1,m_2)	- 96.0383	240.3520
Lognormal	- 100.0472	-

Figures 4-6 show graphs of fitted CDFs of models superimposed on the empirical distribution functions. The graphs suggest that fits using the GLL(m_1,m), GLL(m,m), and lognormal are similar.

5. Conclusion

As demonstrated in examples 1-2, in fitting environmental data, the GLL family of distributions is a good alternative to the lognormal distribution. In particular, on the basis of the values of maximum log-likelihood functions, the GLL(m_1,m_2) seems to be a better probability model for both data sets. Graphs of the CDFs of the models superimposed on the empirical distribution suggest that the GLL(m_1,m_2) generally appears to fit better than the lognormal and other GLL distributions. The use of the family of ALL distributions in fitting environmental data needs to be investigated further.

It is interesting to note that for data of mercury concentration with 115 sample size, the maximum log-likelihood value of GLL(m_1,m_2) distribution is considerably larger than that of the other distributions, especially the lognormal distribution. But for the other data set with a much smaller sample, the improvement in the likelihood function is slight.

The family of probability models may change significantly for the intensity of censoring different types of pollutant, averaging time of interest, different locations, and other factors. Hence, the performance of GLL(m_1,m_2) when incorporating these factors needs to be examined further.

6. References

Bencala, K.E. and Seinfeld, J.H., On frequency distributions of air pollutant concentrations. *Atmospheric Environment, 10,* 941-950, 1976.

Berger, A., Melice, J.L., Demuth, C., Statistical distributions of daily and high atmospheric SO2-concentrations. *Atmospheric Environment, 16* (12), 2863-2877, 1982.

El-Shaarawi, Inferences about the mean from censored water quality data. *Water Resources Research, 25* (4), 685-690, 1989.

Georgopoulos, P.G. and Seinfeld, J.H., Statistical distributions of air pollutant Concentrations. *Environmental Science and Technology, 16* (7), 401A-415A, 1989.

Jakeman, A.J. and Taylor, J.A., A hybrid ATDL-gamma distribution model for predicting urban area source acid gas concentrations. *Atmospheric Environment, 19,* 1959-1967, 1985.

Larsen, R.I., A new mathematical model of air pollutant concentration averaging time and frequency. *Journal of The Air Pollution Control_Association, 19* (1), 24-30, 1969.

Larsen, R.I., An air quality data analysis system for interrelating effects, standards and needed source reductions. *Journal of The Air Pollution Control Association, 23* (11), 933-940, 1973.

Larsen, R.I., An air quality data analysis system for interrelating effects, standards and needed source reductions-part 2. *Journal of The Air* Pollution Control Association, *24* (6), 551-558, 1974.

Lee, L. and Krutchkoff, R.G., Mean and variance of partially-truncated distributions. *Biometrics,* 36, 531-536, 1980.

Mage, D.T., A review of the application of probability models for aerometric data. In Environmetrics 81:SelectedPapers, SIAM-SIMS Conference Series No. 8. Society for Industrial and Applied Mathematics, Philadelphia, 1981.

Millard, S.P. and Deverel, S.J. (1988). Nonparametric statistical methods for comparing two sitesbased on data with multiple nondetect limits. *Water Resources Research, 24* (12), 2087-2098, 1981.

Ott, W.R., Environmental statistics and data analysis. Boca Raton: Lewis, 1995.

Simpson, R.W., Butt, J., Jakeman, A.J. . An averaging time model of SO2 frequency distributions from a single point source. *Atmospheric Environment, 18* (6), 1115-1123, 1984.

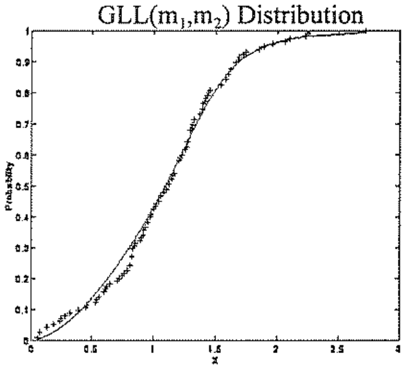
Singh, K.P., A generalized log-logistic regression model for survival analysis: hazard rate characteristics. *Biometrie-Praximetrie, 29*, 63-74, 1989.

Singh, K.P., Bartolucci, A.A., and Burgard, S.L. , Two-step procedure for survival data. *BiometriPraximetrie, 34,* 1-12, 1994.

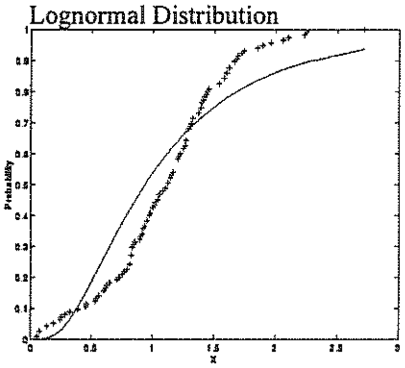
Taylor, J.A., Jakeman, A.J., and Simpson, R.W., Modeling distributions of air pollutant concentratios--I. models. *Atmospheric Environment, 20* (9), 1781-1789, 1986.

Warsono, Singh, K.P., and Bartolucci, A.A., Generalized log-logistic and log-normal models for analysis of environmental pollutant data. Unpublished manuscript, Department of Biostatistics, School of Public Health, The University of Alabama at Birmingham, USA, 1996.

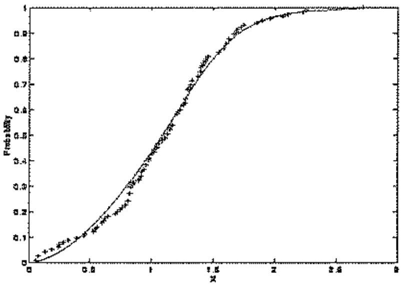
GLL(m_1,m_2) Distribution



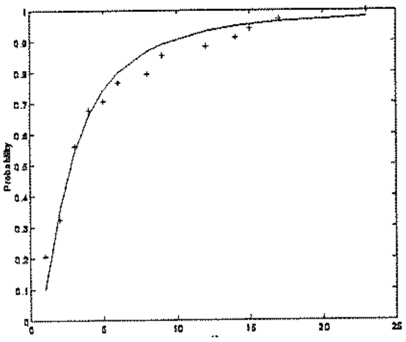
Lognormal Distribution



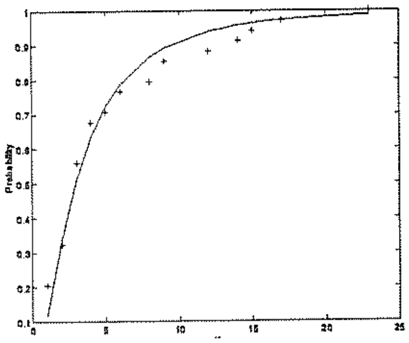
GLL(m,1) Distribution



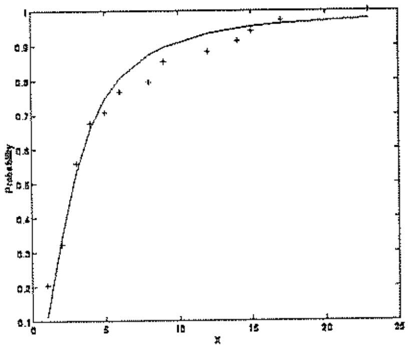
Figures 1-3: Empirical (+) and Fitted GLL and Lognormal Functions - Mercury Concentration Data



GLL(m_1,m_2) Distribution



GLL(m,m) Distribution



Lognormal Distribution

Figures 4-6: Empirical (+) and Fitted GLL and Lognormal Functions - Copper Concentration Data